View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by PhilPapers

ISSN 0147-6882, Scientific and Technical Information Processing, 2011, Vol. 38, No. 2, pp. 119–122. © Allerton Press, Inc., 2011.
Original Russian Text © A.V. Poleev, 2011, published in Nauchno-Technicheskaya Informatsiya, Seriya 1, 2011, No. 6, pp. 19–22.

# Universal Metadata Standard

**A. V. Poleev**
*e-mail: andrejpoleev@yahoo.com*
Received March 14, 2011

**Abstract**—Consciousness is based on the association of notions or a neural network. Similarly, the creation of the next-generation Internet (semantic web) is impossible without attributes that allow the semantic association of documents and their integration into an information context. To achieve these goals, the Universal Metadata Standard (UMS) may serve as a basis for documentography and is functionally required for interpretation of documents by automatic operating systems.

**Keywords:** document, metadata, classification, identification, association, documentography, metagraphy, standard, metabase

**DOI:** 10.3103/S0147688211020122

The goal of scientific knowledge is to embrace the unembraceable. The impossibility of achieving this goal is evident; however, if it is taken as the maxima of scientific and cognitive activities and a guiding star in searching for the truth, then it seems quite reasonable for any individual to come to know more and widen his/her individual circle of knowledge. In essence, most of the time people are involved in organizing the information flow that continuously comes into their brains through sensory organs and receptors, both from the body and from outside. Not only human welfare but also the chances of people for survival are governed by how efficiently such information is ordered to transform the raw material of nerve pulses into true knowledge.

The appearance and development of consciousness refers to the improvement of communication means based on sign information transfer or language. Continuous improvement of communication engineering, overcoming of semantic barriers by the trial-and-error method resulted in standards for the transmission and perception of information, which can be exemplified by the printing industry (polygraphia). Having followed a significant path, polygraphic facilities resulted in microelectronics, which not only improved the quality and widened the scope of true knowledge but also marked the possibility of malicious manipulation with consciousness since the spheres of knowledge production and those document aspects that cannot be directly perceived by an individual are overlooked by readers and viewers, i.e., information recipients. However, they can and should be perceived by information processing machines (computers). The gap that emerged is an intrinsic problem of computer science.

Let us consider an example of how knowledge is organized. The scientific community focuses its attention on the accumulation, verification, and systematization of knowledge shaped as scientific papers. However, any paper is preceded by significant activities that are, as a rule, invisible to the public. The draft of a scientific paper—a laboratory notebook—is a collection of protocols concerning planned experiments and their results. In the ideal case, it should log everything referring to scientific work to be conducted and reflect everything that occurs in a laboratory in chronologic order beginning from goal setting, hypothesis, experimental verification, conclusions, and impressions about everything seen and heard. Formally, the laboratory notebook should describe different format documents, such as photographs, protocol texts, paper texts, lab meetings, references to Internet sources, etc. All these documents should be associated with each other, provided with comments, and should be accessible for reviewing and cataloging. For example, experiments are ideas referring to various themes that can follow in chronological order after each other, including theoretical research into a problem and collection of the relevant data; writing a book or a paper based on the work done; and the planning of thematically different experiments. In this connection, this thematic diversity should be shown in the list of themes, as well as in the possibility of extracting similar (allied) information by means of thematic tags and location tags.

As a MacBook computer user I can accumulate and thematically unite different documents. However,

additional software is required for their description and visualization. File Maker in part satisfies the needs for systematization and description since at this stage an appropriate browsing panel is lacking and the opportunity to open and use the documents inside the program while avoiding additional ones that in the ideal case should be built in as options rather than being scattered in different places, e.g., a web editor, web browser, photoshop, file maker, pdf reader, video or photo visualizer, text editor, etc.

In the context of the variety of the document base of scientific consciousness and knowledge, data documentation and systematization becomes of primary importance. Documents are commonly classified by alphabet, date, theme, project, format, or location (local folder, internet address). For their identification a date, number, and name are used. For example, images have jpg, gif, png, and psd file formats and texts can be pdf, doc, or txt. The document format is its identification tag that is required to recognize it in operation systems and in program initiation (processing). However, its systematic description is still not present in each format necessary and sufficient for integration and transfer to other descriptive systems (e.g., when copying from an electronic library to a personal computer). Any document reflects real subjects and events, describes them, and comprises certain attributes. However, a photograph does not retain information about an object's size, origin, history, and goal. All this, in the ideal case, should be included into a metainformation supplement to the document, at least in the form of references. An increasing number of documents and formats are not followed by the improvement of engineering possibilities of their perception and systematization. Instead, descriptive systems (doi, ISBN, URN, PURL, ISNI, etc.) and aliases are multiplied. For example, a journal paper, as a rule, in html or pdf format in the NCBI/NLM descriptive system, receives a number (PUBMED ID), also, an abstract with a publication date, journal title, authors' names, and keywords are added. However, these data should be added directly to the document as a supplement or extension in order to have a chance to order the document when moving it to other descriptive systems (e.g., in translating it to another language or using it in another database). The history of this movement (e.g., when copying it from an electronic library) should be reflected in the document. To achieve this goal, a universal standard for all document types should be created. Also, it is necessary to fix the options that are present in each format, how they will be filled in or modified, and which items should remain unchanged. I propose the following clear options for the metadata description of documents:

A unique name[1]

Format

Date

Classification system used

Identity number

Language[2]

Position and location

Creator, origin, or source.

It can be conceived that instructions on the manufacture of an atomic weapon or pornographic documents cannot be accessible to everyone. Therefore, an accessibility gradation should be introduced to limit the access to documents.

If a document undergoes modification (transfer to another descriptive system, size or name change) then primary metadata should be maintained and changes should be written automatically or manually: a synonym should be added during renaming; in another descriptive system a new internal designation and ID number should be added; when transposed, a new Internet address or geographic position should be given, etc.

For each attribute of a universal metadata standard one should determine an option form, define it, and formally describe it. The content of each option should correspond to the rules based on which the catalog of permissible systematic designations could be complied. For example, document authorship should

[1] A systematic designation is a sequence of symbols based on which an object is identified and a correspondence is set between its perception through organs of the senses (sensory representation) and linguistic interpretation of this perception. A systematic name should have attributes to allow the object to be referred to a certain designation class, as well as containing a required supplement that is sufficient for unambiguous identification among allied names and designations. By way of example, in a narrow circle of individuals the name Andrew is sufficient, while in a group comprising individuals with the same name a family name is necessary. At a global scale, a name, birth date, and birth place are sufficient for identity. A systematic designation to define a person can consist of two or three names, a sequence of figures, and a geographic definer. In a similar way, the systematic designation of an organization can contain a name, foundation date and place, and date when its activities ended. The answers to three questions: who or what, when, and where, are sufficient for identification in other cases as well.
The words catalog, nomenclature, classification, and register are largely synonymic and are lists of designations that are united in allied groups that in turn can be united based on certain criteria. The order of unification can vary depending on the chosen criteria. The names of persons can be grouped by alphabet based on the birth date or the place of their prototypes. In the dynamic of space categorization a systematic designation remains a constant and crystallization point that allow semantic association, searching, and ascertaining the relations between designations, concepts, definitions, and categories.

[2] A language means sign systems of natural languages that are principally descriptive and indicative in contrast to computer programs that are derivatives of natural languages and have the directive nature of algorithms, i.e., instructions for automatic operators.

be unambiguous and based on an authors' list. A document's origin should be based on a list of organizations. A document type (text, figure, photograph, video, or sound) should be followed by the description (summary) and typological attributes that are characteristic for each document type. Every document should comprise a list of objects or events that are reflected in it (biological species, astronomic object, person or group of individuals, organization, scientific paper, etc.). Presently, the classification basis of such a list exists (Encyclopedia of Life, International Plant Names Index, Catalogue of Astronomical Objects, ICD, etc.) and should be used in the UMS.

What happens in reality? Let us consider a bright example. The extraction of metadata for the document octology.pdf found at the http://www.enzymes.at/download/octology.pdf address gives the following result:

CreateDate = 2011:03:01 16:35:22Z
Title = octology
PageCount = 76
FileSize = 11 MB
Author = Max Madman
MIMEType = application/pdf
PDFVersion = 1.4
FileType = PDF
Creator = Pages
ModifyDate = 2011:03:31 16:35:22Z
PDFVersion (1) = 1.3
Producer = Mac OS X 10.5.2 Quartz PDFContext.

The description is clearly senseless: the pdf format is indicated six times; the document's creator and author are unknown; the dates of creation and modification of the document coincide and tell nothing about the time of its appearance. Only the indication of pages and document size make sense. Metadata on the figures contained in the text (if any) are totally lost in the pdf format. Document publication on the Researchgate.net portal is accompanied by the DOI address: details/Octology. It is unclear what this means, since the verification of the address in the descriptive system does not provide any results. Although the journal with the document is included in the NCBI/NLM database, the data about this document is still absent in the PUBMED electronic library. Information cannot be included manually, since everything is automated and this function cannot be executed. The system fails because the data about the document titled Octology is absent in PUBMED/NCBI/NLM but present in the associated OCLC/WorldCat descriptive system.

The metadata of a document chosen randomly from the PUBMED library look even more absurd: Palesch D., Sienczyk M., Oleksyszyn J., Reich M., Wieczerzak E., Boehm B.O., Burster T., Was the serine protease cathepsin G discovered by S. G. Hedin in 1903 in bovine spleen & Acta Biochim Pol. 2011 Mar 7, PMID : 21383996 (as in the previous case, extraction can be executed by means of ServerSniff software available on the Internet at http://serversniff.net/file-info.php).

Summing up, it is reasonable that programmers, terminologists, ISO, and the knowledge industry would develop a logically verified system of metainformation provision for general use. The system assumes that document production won't become an end in itself and would acquire a reliable base that makes it possible to efficiently adopt and organize knowledge at a new social and engineering stage. In parallel, it is advisable to supplement programs with a module allowing the visualization and edition of metadata, as well introducing universal programs for all types of documents (metadata editors).

The general thematic idea of this paper is the creation of a complex of semantic standards, with UMS being one of them. Nikola Tesla who introduced the technical principles of the Internet at the turn of the 19th and 20th centuries pursued the idea of the disappearance of the borders impeding communication and cognition. Today, the Internet creates a virtual reality based on which reality, consciousness, and society are constructed. Some threats context should be mentioned in this respect. By way of example, the semantic content of ontology[3] as one of the focus notions of the third-generation Internet is intentionally distorted for

[3] Since the being of objects is manifested in actions then the description of interactions in this set of objects gives us an idea about the domain under study. An ontological scheme is a formalized description of associations and interactions between objects in a certain set of objects. A scientific domain including the set of objects and phenomena under study, study and description methods, hypotheses, and theories can serve an example of the application of ontological schemes. Another example is an enterprise, which includes equipment (means of production), technological description of production (methods of production), behavior rules for employees (instructions on enterprise management), and other conditions for its functioning. In the center of an ontological scheme there is the description of objects including a name, address, and attributes (qualities and properties of manifestation). Any description is based on the systematization to refer the object under study to a group of objects of the given ontological scheme. Here the object attributes can be of the general nature of systematic categories based on which the entire set of objects is divided into subgroups. For instance, in a set of things, some of them can be spherical in shape, differ in color, etc. Therefore, the differentiation of objects occurs by systematization based on individual attributes and categorization is a recursive procedure that distinguishes the necessary and sufficient object attributes to systematize them and distribute them inside the given set of objects. However, ontological schemes cannot only describe reality but also actively affect objects and govern their behavior status by setting the interaction rules. A subjective factor of ontological schemes manifests itself in state management based on incomplete, distorted, or inadequate description of objects, i.e., people, social groups, and their relationships, as well as excluding from consideration ontological schemes of a more general nature (environment, biosphere, cosmology, philosophy, etc.). No wonder that the people in these ontologies are still considered as consumables to be treated as things or livestock.

ideological purposes. Thus, in business, ontologies are logical schemes that are intended to manipulate consciousness, to put certain stereotypes into people's heads, and to promote group interests. Ontological schemes written in an artificial language inaccessible to a wide public are intended to provide the latent control of a narrow circle of individuals over society. In this context, the semantic Internet can become a tool for totalitarian manage of a global scale. It is clear that the seizure of power may occur secretly and the proper totalitarian process will be beyond juridical regulation. In order to exclude the malicious usage of Internet technology it is necessary to take measures in advance. The universal standards proposed in this paper allow the above-described scenario to be avoided and Internet regulation to be accessible to its users.

Uninformed readers can bridge a gap in their knowledge by becoming familiar with the sources listed in the references. References [1–17] concern metagraphy, [18–21] are related to metadata in arts, literature, and philosophy, and [22–25] concern engineering means for the organization of scientific literature.

## REFERENCES


1. Semantic Internet. http://semanticweb.org/
2. Bergman, M.K., A Timeline of Information History. http://mkbergman.com/temp-exhibit/
3. Handbook of Metadata. Semantics and Ontologies, 2012, World Scientific Publishing. http://www.worldscibooks.com/compsci/7077.html
4. Health, T. and Bizer, Ch., Linked Data: Evolving the Web into a Global Data Space, *Synthesis Lectures on the Semantic Web: Theory and Technology*, 2011, vol. 1, no. 1.
5. Bioontologies. http://obofoundry.org/
6. Music Ontology. http://musicontology.com/
7. Ontologies for Electronic Government. http://www.oegov.org/
8. Building Ontologies. http://en.wikipedia.org/wiki/Ontology_engineering
9. Document Metadata. http://en.wikipedia.org/wiki/Metadata_standards
10. The List of File Formats. http://en.wikipedia.org/wiki/List_of_file_formats
11. Adobe XMP. http://www.adobe.com/products/xmp/
12. Metadata Standards. http://en.wikipedia.org/wiki/Metadata_standards
13. eXtended Metadata Registry (XMDR) Project. https://xmdr.org/overview.html.
14. Gill, T., Gilliland, A.J., Whalen, M., and Woodley, M.S., Introduction to Metadata, ed. by Murtha Baca, Getty Publications, 2008. http://www.getty.edu/ research/publications/electronic_publications/introm-etadata/index.html
15. RIP Keywords Meta Tag. http://www.seohosting.com/blog/articles/rip-keywords-meta-tag/
16. Miller, S.J., *Metadata and Cataloging Online Resources*, 2010.
17. Greenberg, J., Metadata and Digital Information, Bates, M.J., Maac, M.N., and Drake, M. Eds., in *Encyclopedia of Library and Information Science*, New York: Marcel Dekker, 2009.
18. *Mark Amerika META/DATA A Digital Poetics*, MIT Press, 2007.
19. Metaexhibition 2010. http://www.metaexhibition.ca/
20. Tsoukas, H. and Knudsen, Ch., *The Handbook of Organization Theory: Meta-Theoretical Perspectives*, Oxford: Un-ty Press, 2003.
21. *Metadata Symposium, Academy of Motion Picture Arts and Sciences.*
22. Laboratory Notebook. http://openwetware.org/wiki/Lab_Notebook
23. Amber Dance. How to Choose Your Lab's Next Electronic Lab Notebook, *The Scientist*, 2010, vol. 24, no. 5, p.71.
24. EndNote. http://www.endnote.com/
25. Mekentosj Papers. http://www.mekentosj.com/papers/.